ロボティクスと言語における統計的分節化

持橋大地
統計数理研究所 数理・推論研究系
daichi@ism.ac.jp

RSJ2016
2016-9-8 (木)@山形大学

自己紹介

- 奈良先端大・自然言語処理学講座 博士後期課程修了
- 2011～：統計数理研究所 准教授
- 専門：統計的自然言語処理/計算言語学
- ひとこと：
  ロボティクスの皆さん、IBIS 2016 (情報理論的学習理論ワークショップ; 機械学習の国内最大の会議) にぜひ来て下さい! (2016/11/16-19 @京大)
  - ロボティクスの人の参加が少ない
  - 今後、ロボティクスは機械学習的にも極めて重要

分節化とは

- 時系列を、意味のある単位に区切ること
- ロボティクスの場合：動作
- 自然言語の場合：単語 (最小限の単位として)
  - 人間は、教えられなくてもこれらを認識できる

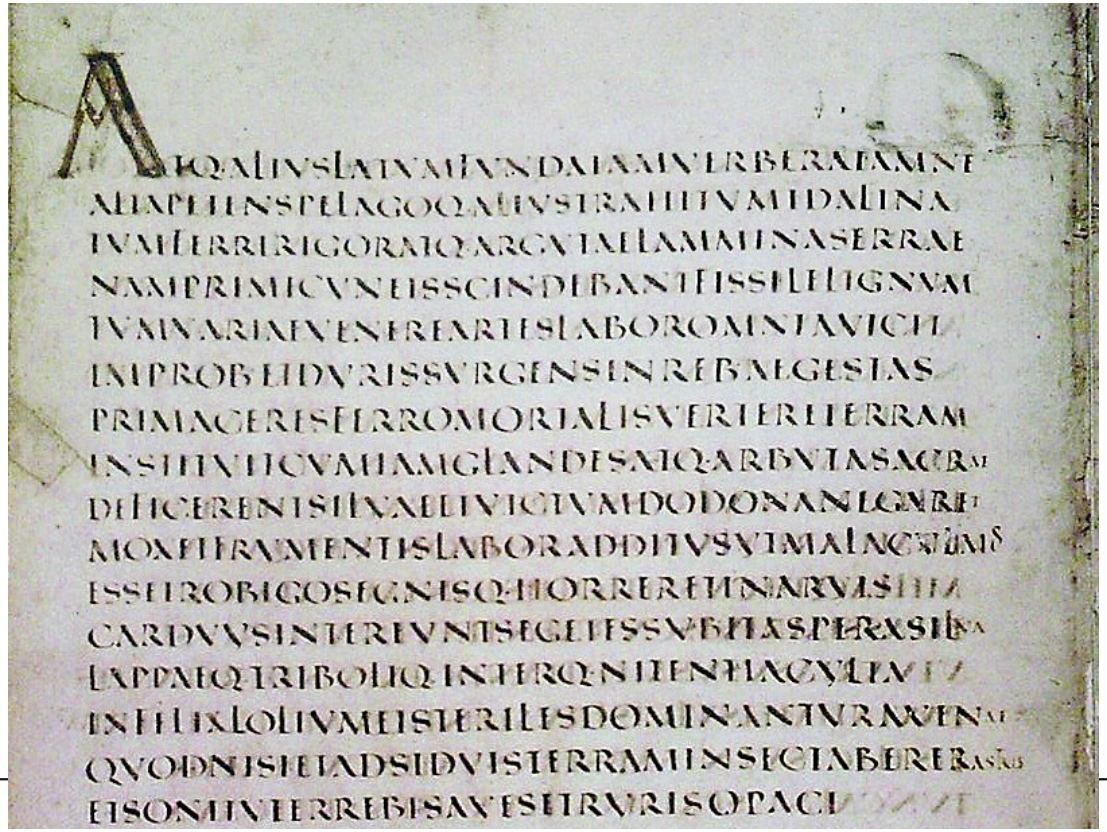
言語の場合

- 中国語
  - 話者12億人なので、重要な研究
  - Weibo (中国版Twitterクローン) の一部

今个学期第一次来图书馆呀～嘻嘻～感觉好好呀～稳紧
写论文既资料呀～收获唔多，睇来要上网“刮”料拉
活动精彩照片集锦8-行进途中，璀璨夜色
到底这次是不是真的要走啦！！！tnnd敢不敢不要再
变来变去了[愤怒]

言語の場合 (2)

- ラテン語 (*Scripta continua*)
  - 英語も最初は、単語間スペース抜きで書かれていた



ヴェルギリウスの文、
AD141年前後

従来の形態素解析 (Chasen, MeCab, JUMAN..)

```
# S-ID:950117245-006 KNP:99/12/27
* 0 5D
一方 いっぽう * 接続詞 * * *
、 、 * 特殊 読点 * *
* 1 5D
震度 しんど * 名詞 普通名詞 * *
は は * 助詞 副助詞 * *
* 2 3D
揺れ ゆれ * 名詞 普通名詞 * *
の の * 助詞 接続助詞 * *
* 3 4D
強弱 きょうじゃく * 名詞 普通名詞 * *
```

毎日新聞
1995年度記事
から38,400文
(京大コーパス)
の例

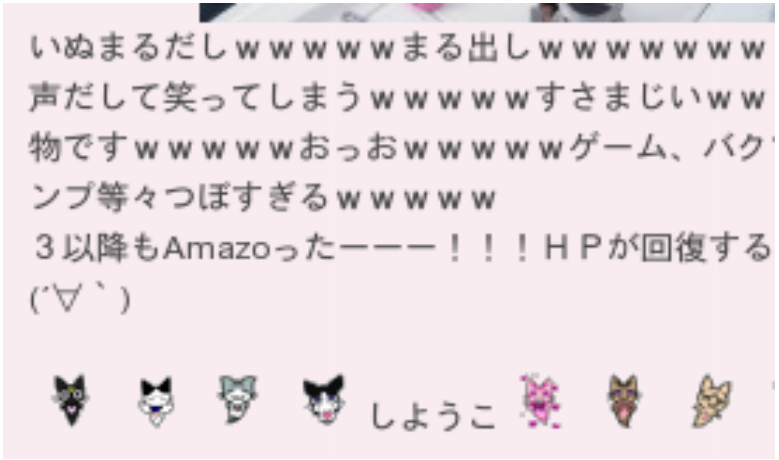
- 膨大な人手で作成した教師(正解)データ
- 話し言葉の「正解」？古文？未知の言語?
  - |女御|更衣|あ|また|さ|ぶら|ひ|た|ま|ひける|中|に|、|
    …

教師あり学習の限界

- 従来の新聞記事コーパスでは扱えない言語データが増えている (口語, 新語, 新表現, ...)

Twitter

Blog



– 全部人手で辞書登録...？(追いつかない!)

話し言葉の解析

CSJ話し言葉コーパスの一部

何て言うんでしょうねその当時はそう思われていたっていうことを全部ドラマにしちゃうっていうところそういうところとかが面白くて凄く見てたんですけど最初の方は凄くまともで緋色の研究とか四人の署名とかそういうのから始まってたんですけどそれはもうとにかく原作に沿ってこういうこれこれホームズってこれっていうのが見たくて私はずっと見てます見てましたで凄くこれは何かＮＨＫが放送されてる時から...

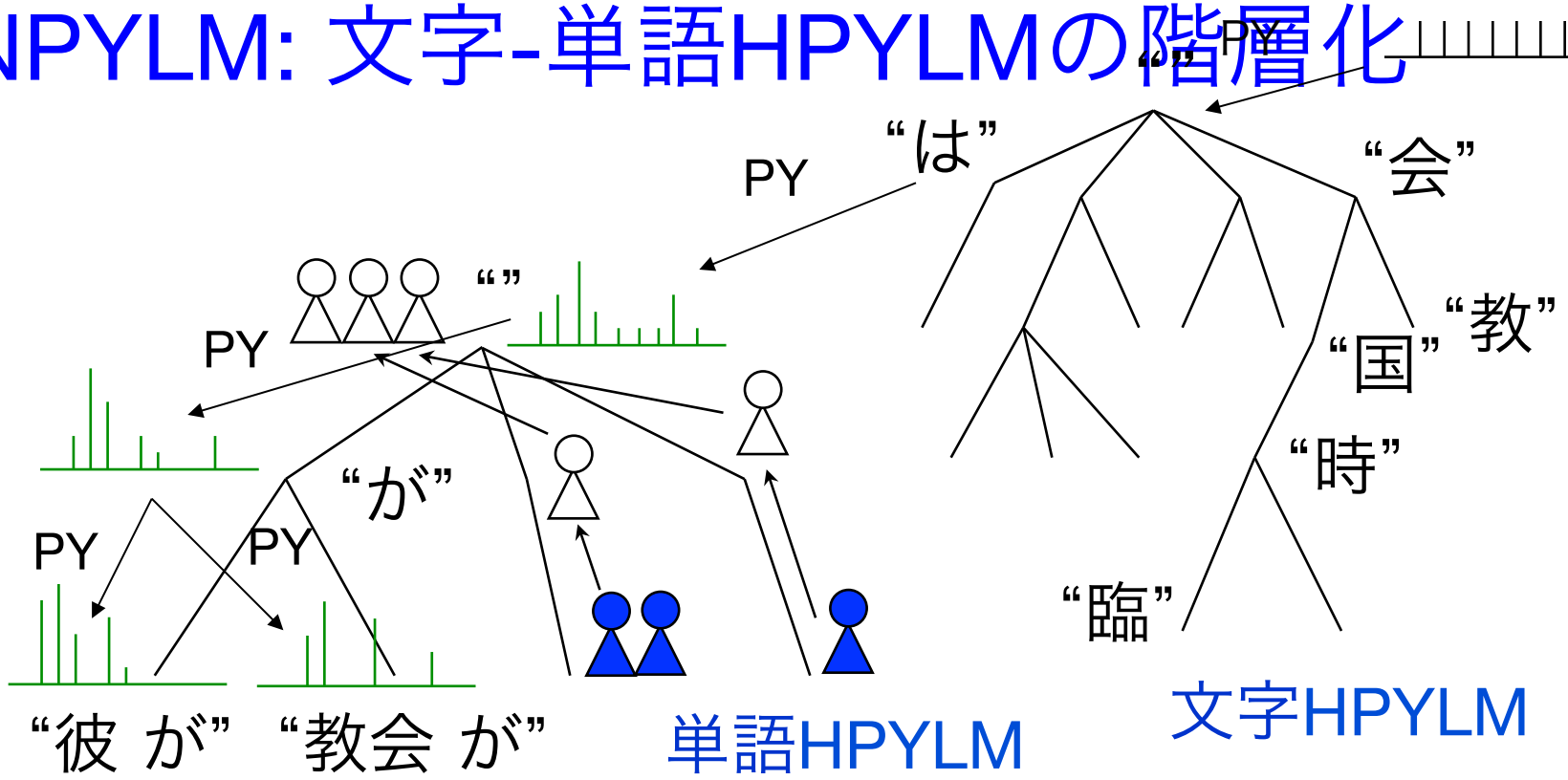
- 多数の口語表現 (“そいで”,“ってさぁ”,“ちゃって”...)
  - 無数のバリエーションが存在
- 自然な話し言葉の音声認識や音声科学のために、長期的にきわめて重要

教師なし形態素解析 (持橋+, ACL2009)

- 生の文字列だけから, 階層ベイズで「単語」を学習
  - モデル: NPYLM (Nested Pitman-Yor LM)

1 神戸では異人館　街の 二十棟 が破損した 。
2 神戸 では 異人館 街の 二十棟 が破損した 。
10 神戸 では 異人館 街の 二十棟 が破損した 。
50 神戸 で は異人 館 街 の 二 十 棟 が 破損 し た 。
100 神戸 で は 異 人館 街 の 二 十 棟 が 破損 し た 。
200 神戸 で は 異人館 街 の 二 十 棟 が 破損 し た 。

NPYLM: 文字-単語HPYLMの階層化



- 単語nグラム-文字nグラムの埋め込み言語モデル
  - つまり、階層Markovモデル
- 階層ベイズモデル+Gibbsサンプリング

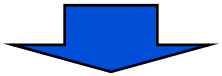
日本語話し言葉コーパス (国立国語研究所)

うーんうんなってしまうところでしょうねへーあーでもいいいいことですよねうーん
うーん自分にも凄くプラスになりますものねそうですねふーん羨ましいです何かうーん精神的にもう子供達に何かこう支えられるようなうーものってやっぱりあるんですよやってるとうーんうーんうーん
うーん長くやってればそんなものがうんうんそうでしょうねたくさんやっぱりありますねうんうーんなるほど…

うーん うん なって しまう ところ でしょう ね へー あー でも いい いい こと ですよねうーん
うーん 自分 に も 凄く プラス に なり ます もの ね そう です ね ふーん 羨ましい です 何か うーん 精神的にもう 子供達に何か こう 支えられる ような うー もの って やっぱり ある んです よ や って る と うーん うーん うーん うーん 長く や って れば そんな もの が うん うん そう でしょう ね たくさん やっぱり あり ます ね うん うーんなるほど…

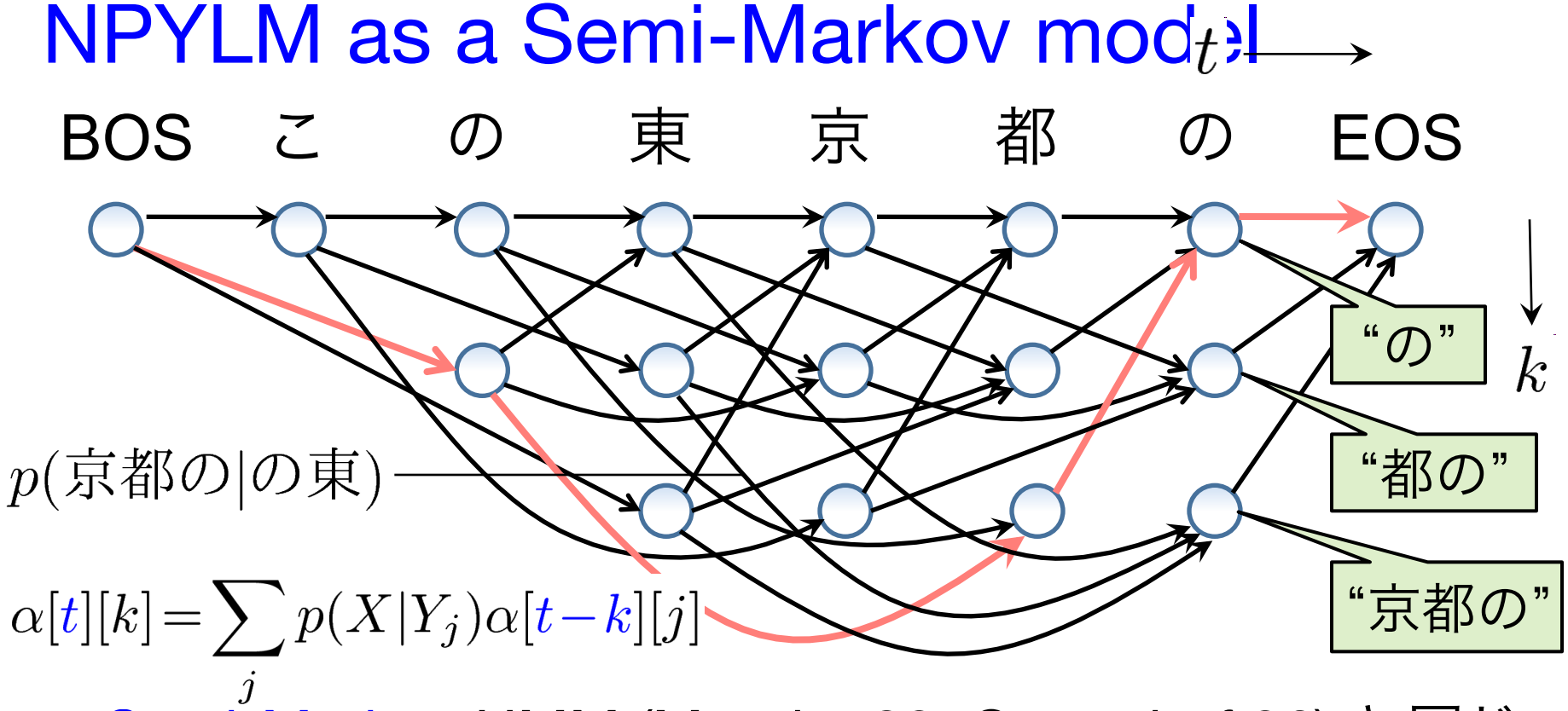
“Alice in Wonderland”の解析

first,shedreamedoflittlealiceherself,andonceagainthetinyhandswereclaspedup onherknee,andthebrighteagereyeswerelookingupintohersshecouldhearthevery tonesofhervoice,andseethatqueerlittletossofherheadtokeepbackthewanderingh airthatwouldalwaysgetintohereyesandstillasshelistened,orseemedtolisten,thew holeplacearoundherbecamealivethestrangecreaturesofherlittlesister'sdream.the longgrassrustledatherfeetasthewhiterabbithurriedbythefrightenedmousesplashe dhiswaythroughtheneighbouringpoolshecouldheartherattleoftheteacups…



first, she dream ed of little alice herself ,and once again the tiny hand s were clasped upon her knee ,and the bright eager eyes were looking up into hers -- shecould hearthe very tone s of her voice , and see that queer little toss of herhead to keep back the wandering hair that would always get into hereyes -- and still as she listened , or seemed to listen , thewhole place a round her became alive the strange creatures of her little sister 'sdream. thelong grass rustled ather feet as thewhitera bbit hurried by -- the frightened mouse splashed his way through the neighbour ing pool -- shecould hearthe rattle ofthe tea cups …

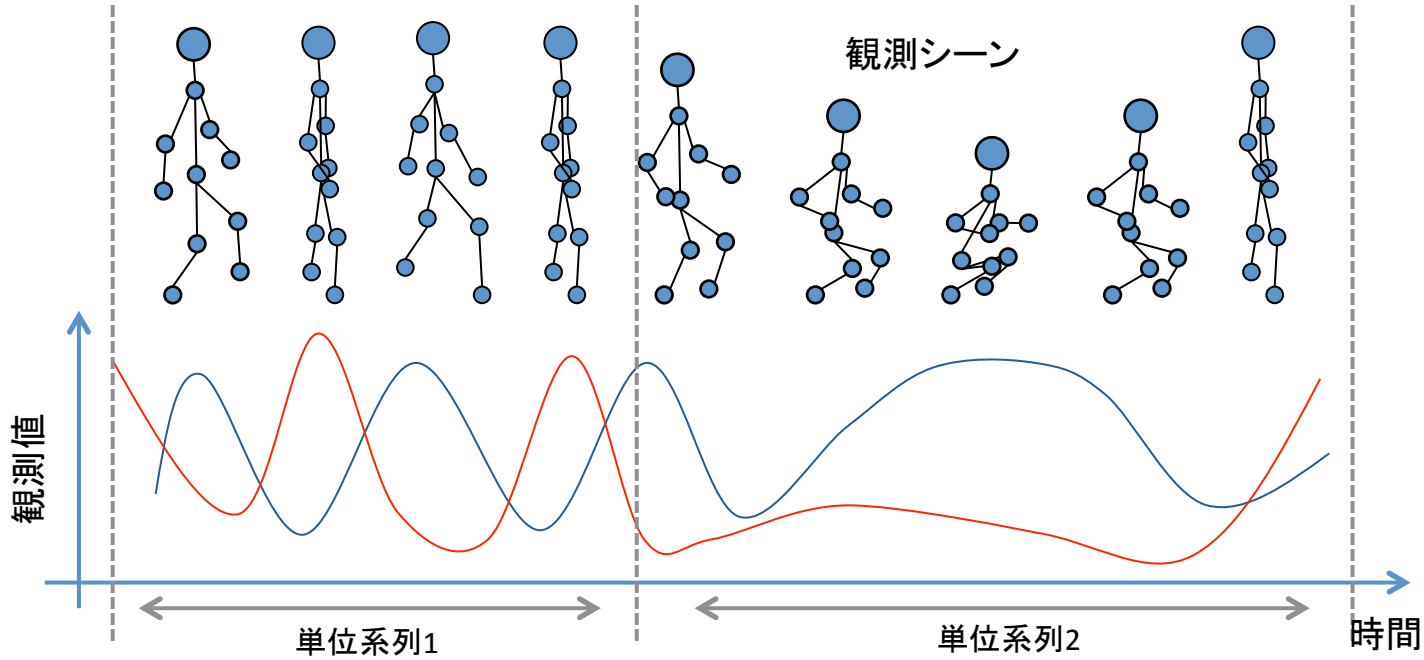
NPYLM as a Semi-Markov model



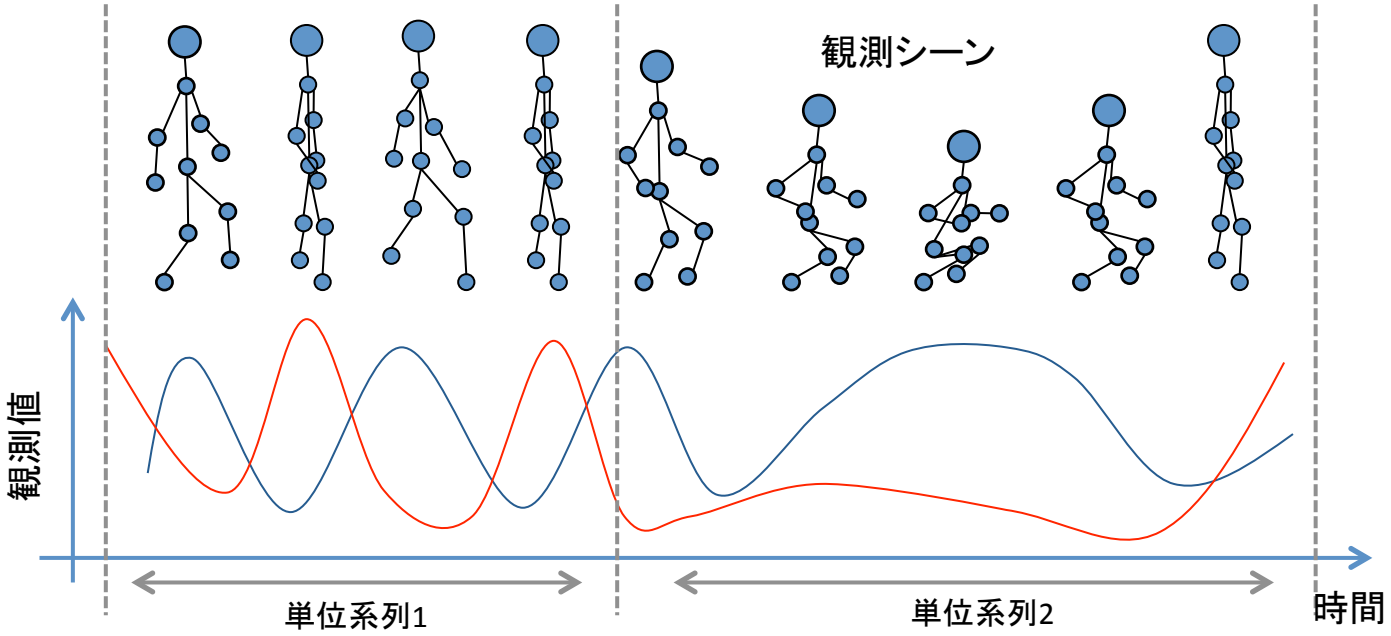
- Semi-Markov HMM (Murphy 02, Ostendorf 96) と同じ
  - 正しいパスが何かは未知 (HMMと同じ)
- 動的計画法によって、確率の高いパスをサンプリング

ロボティクスへの応用

- 関節角の時系列データから、教師なしで「動作」を認識したい
  - 中村さん・長井先生との共同研究 (2015-16)



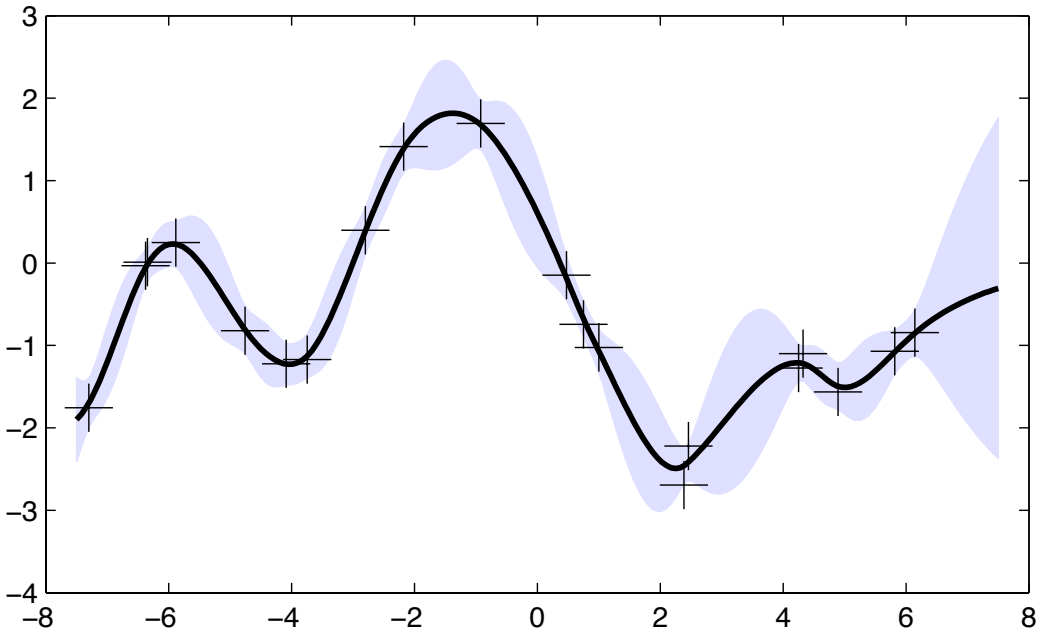
ロボティクスへの応用



- 出力が連続 → “軌跡”を出力するような確率モデルが必要
- ガウス過程によって実現

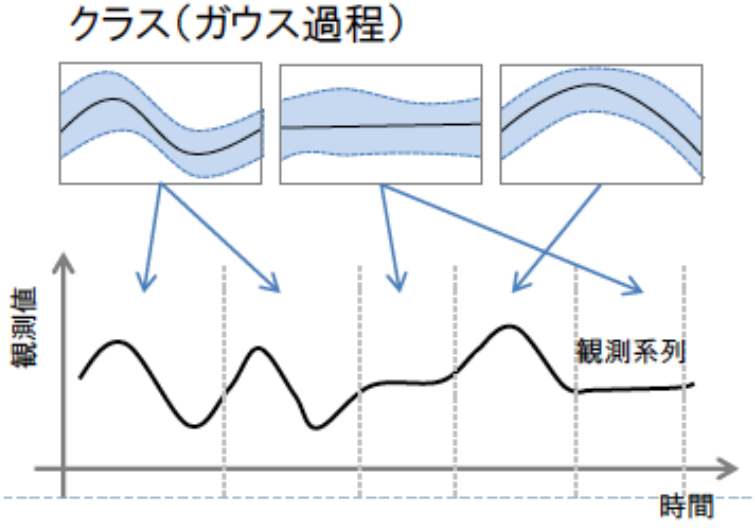
ガウス過程とは

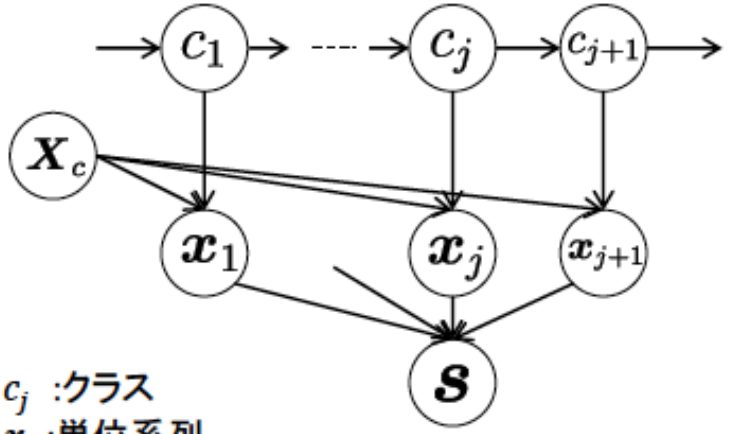
- ランダムな関数を生成する確率モデル
  - 軌跡全体が、無限次元ガウス分布の一点に対応
- データが与えられれば、関数の事後分布がガウス分布で得られる



提案モデル

- 時系列データはガウス過程を出力分布とする隠れセミマルコフモデル(HSMM)によって生成されると仮定
  - 状態(動作クラス)の決定: $c_j \sim P(c \mid c_{j-1})$
  - 時系列データの生成: $\boldsymbol{x}_j \sim GP(\boldsymbol{x} \mid \boldsymbol{X}_{c_j})$
- HSMMとGPのパラメータ推定することで，系列データの分節化・分類が可能



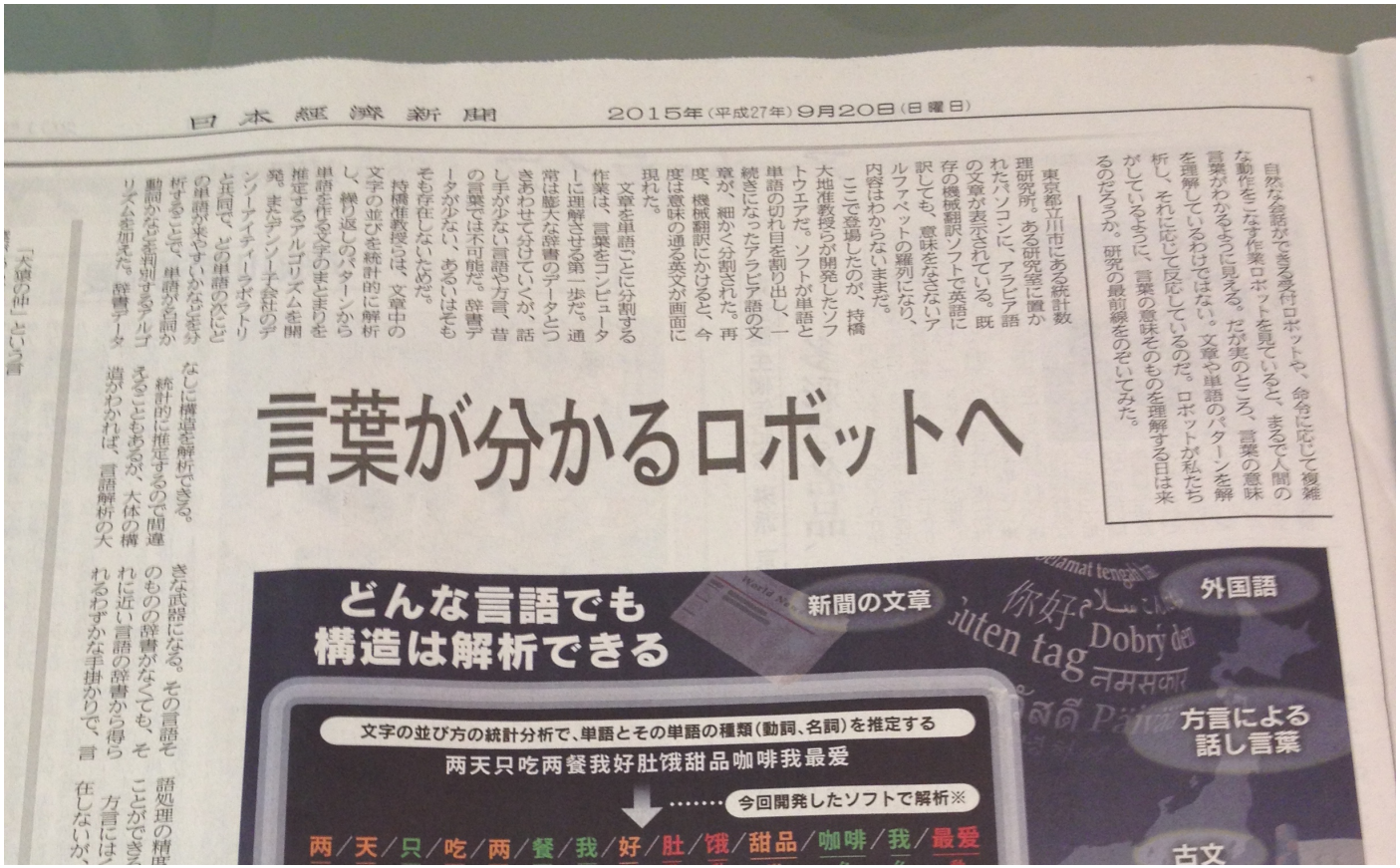


c_j :クラス
$\boldsymbol{x}_j$:単位系列
$\boldsymbol{X}_c$:クラスcのガウス過程のパラメータ
$\boldsymbol{S}$:観測系列

注意

- ここまでは、分割された“単語”に状態を考えてこなかった
  - HMMではなく、Semi-Markovモデルだった
- 原理的には、HMM化することで言語でも、“品詞”を同時に求めることが可能→Hidden Semi-Markovモデル
  - 内海・塚原・持橋 (ACL 2015)で提案

2015/9/20,日経新聞日曜版科学面



日本経済新聞　2015年(平成27年)9月20日(日曜日)

自然な会話ができる受付ロボットや、参会に応じて複雑な動作をこなす作業ロボットを見ていると、まるで人間のように見える。だが実のところ、言葉の意味を理解しているわけではない。文章や単語のパターンを解析し、それに応じて反応しているのだ。ロボットが私たちがしているように、言葉の意味そのものを理解する日は来るのだろうか。研究の最前線をのぞいてみた。

東京都立川市にある統計数理研究所。ある研究室に置かれたパソコンに、アラビア語の文章が表示されている。既存の機械翻訳ソフトで英語に訳しても、意味をなさないアルファベットの羅列になり、内容はわからないままだ。

ここで登場したのが、持橋大地准教授らが開発したソフトウエアだ。ソフトが単語と単語の切れ目を割り出し、一続きになったアラビア語の文章が、細かく分割された。再度、機械翻訳にかけると、今度は意味の通る英文が画面に現れた。

文章を単語ごとに分割する作業は、言葉をコンピュータに理解させる第一歩だ。通常は膨大な辞書のデータとつきあわせて分けていくが、話し手が少ない言語や方言、昔の言葉では不可能だ。辞書データが少ない、あるいはそもそも存在しないためだ。

持橋准教授らは、文章中の文字の並びを統計的に解析し、繰り返しのパターンから単語を作る文字のまとまりを推定するアルゴリズムを開発。また、デンソー子会社のデンソーアイティーラボラトリと共同で、どの単語の次にどの単語が来やすいかなどを分析することで、単語が名詞か動詞かなどを判別するアルゴリズムを加えた。辞書データなしに構造を解析できる。

統計的に推定するので間違えることもあるが、大体の構造がわかれば、言語解析の大きな武器になる。その言語そのものの辞書がなくても、それに近い言語の辞書から得られるわずかな手掛かりで、言語処理の精度

ことができる。方言には存在しないが、

「犬猿の仲」という言

言葉が分かるロボットへ

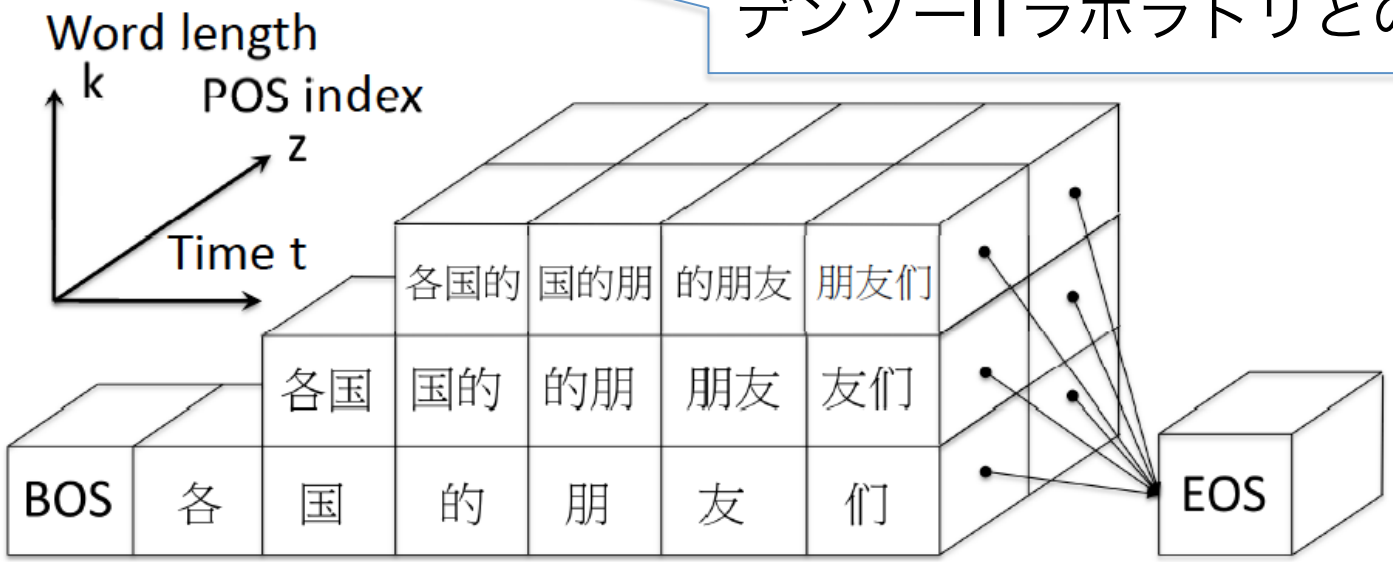
(7/27の朝刊科学技術面でも紹介)

- デンソーITラボラトリ(渋谷)との共同研究

隠れた「品詞」との同時学習 (ACL 2015)

- 単語だけでなく、その“品詞”も知りたい→同時学習
  - MCMCの前向き確率は $\alpha[t][k][z]$ (k:単語長, z:品詞)
  - Amazon EC2と並列化を使った膨大な計算量

デンソーITラボラトリとの共同研究



各国的朋友们
“friends of each country”

実験結果

- 日本語・中国語・タイ語の教師なし単語分割において世界最高精度 (可能な最高値は84.8%)
  - 中国語は12億人の話者がいるため、重要な研究

Dataset	PYHSMM	NPY	BE	HMM^2
Kyoto	**71.5**	62.1	71.3	NA
BCCWJ	**70.5**	NA	NA	NA
MSR	**82.9**	80.2	78.2	81.7
CITYU	**82.6***	82.4	78.7	NA
PKU	**81.6**	NA	80.8	81.1
BEST	**82.1**	NA	**82.1**	NA

NPY: Nested Pitman-Yor (Mochihashi+ 2009)

BE: Branching Entropy (Zhikov+ 2010)

HMM^2: Char and word HMMs (Chen+ 2014)

前向き確率の計算

$$\alpha[t][k][c] = p(\boldsymbol{x}_{t-k:k}|\boldsymbol{X}_c)\sum_{k'}\sum_{c'} p(c|c')\alpha[t-k][k'][c']$$

- 出力確率： $p(\boldsymbol{x}_{t-k:k}|\boldsymbol{X}_c) \propto \underbrace{GP(\boldsymbol{x}_{t-k:k}|\boldsymbol{X}_c)}_{\text{ガウス過程}}\underbrace{p(k)}_{\text{ポアソン分布}}$
- 遷移確率：ディリクレ分布を事前分布とした多項分布

$$p(c|c') = \frac{N_{cc'} + \phi}{N_c + C\phi}$$

- 前向き確率：

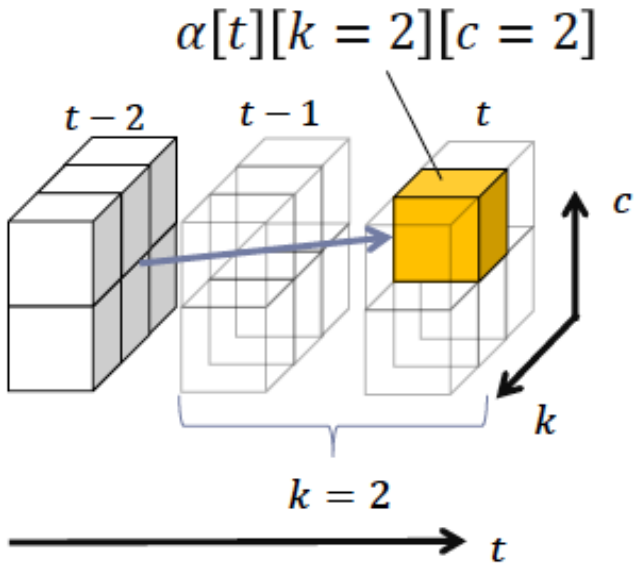
$$\alpha[t][k][c] = GP(\boldsymbol{s}_{t-k:k}|\boldsymbol{X}_c)p(k)\sum_{k'}\sum_{c'}\frac{N_{cc'} + \phi}{N_c + C\phi}\alpha[t-k][k'][c']$$

Forward filtering

- ある時刻tを終点とした長さがk，クラスがcとなる確率を計算
- 分節は$t-k$を終点とする分節から遷移する可能性がある
  ➡ この可能性を周辺化することで再帰的に計算可能

$$\alpha[t][k][c] = p(\boldsymbol{x}_{t-k:k}|\boldsymbol{X}_c)\sum_{k'}\sum_{c'} p(c|c')\alpha[t-k][k'][c']$$

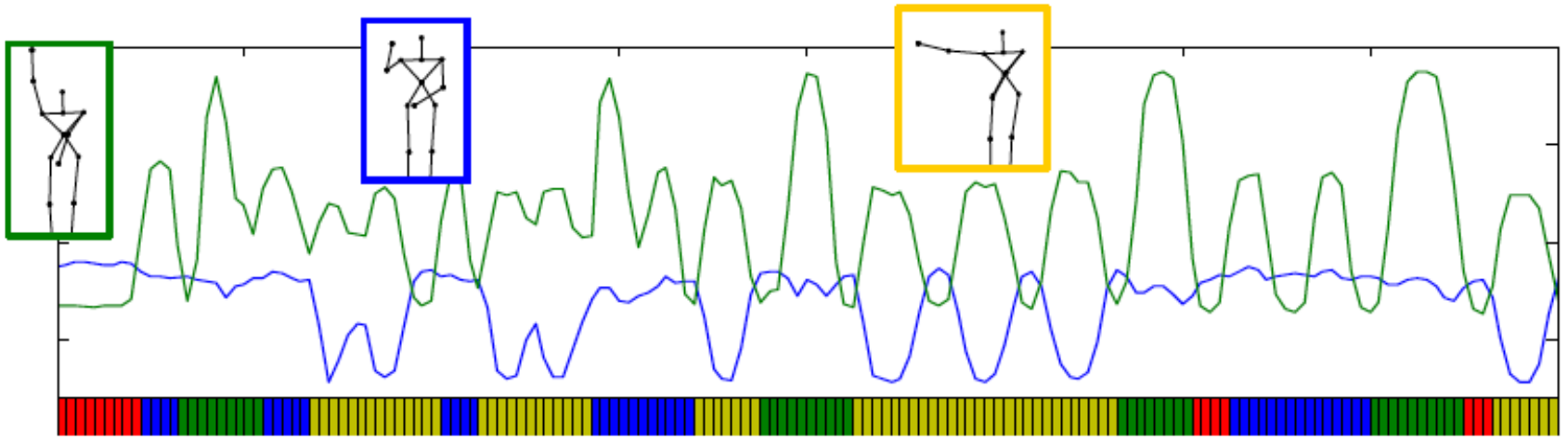
$$\begin{aligned}\alpha[t][2][2] = p(\boldsymbol{x}_{t-2:t}|\boldsymbol{X_2}) \times \\ \{\, & p(2|1)\alpha[t-2][1][1] \\ & +p(2|1)\alpha[t-2][2][1] \\ & +p(2|1)\alpha[t-2][3][1] \\ & +p(2|2)\alpha[t-2][1][2] \\ & +p(2|2)\alpha[t-2][2][2] \\ & +p(2|2)\alpha[t-2][3][2]\}\end{aligned}$$



- $\alpha[0][*][*]$から動的計画法で前向きに計算

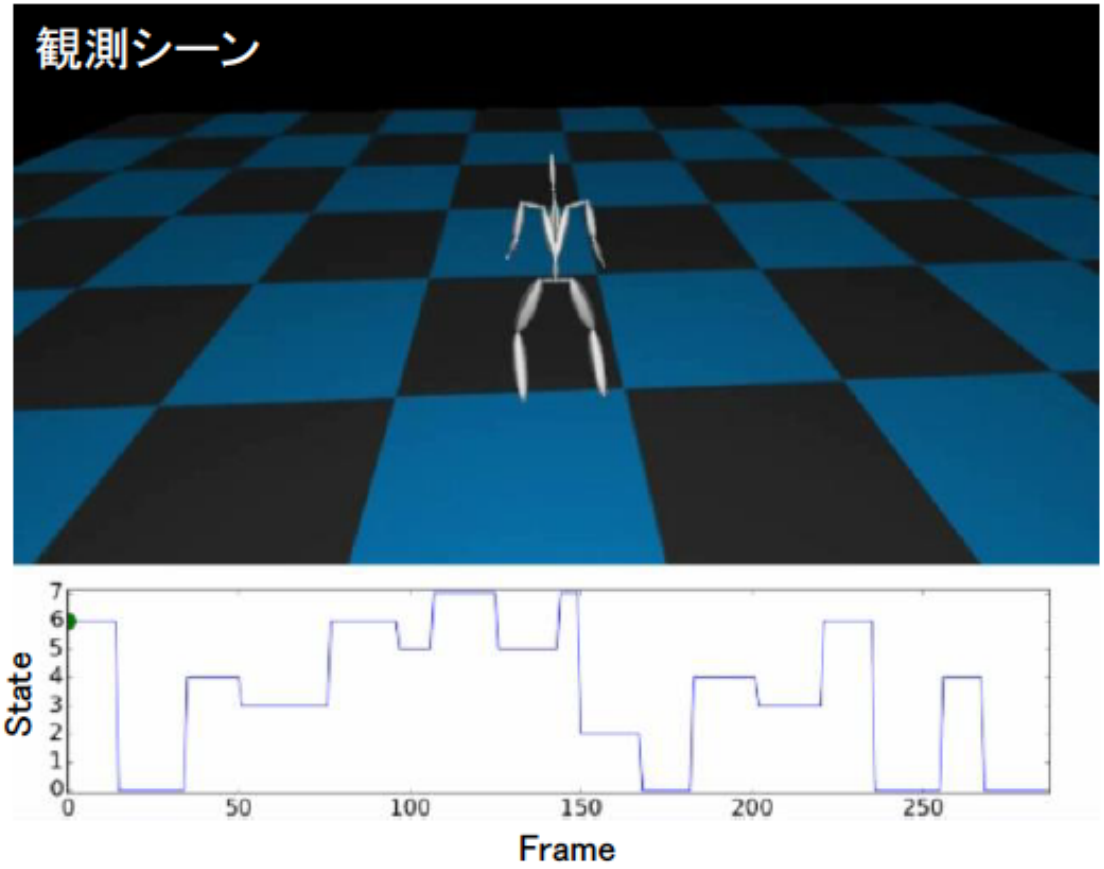
ロボティクスへの応用

- 出力がガウス過程となる隠れsemi-Markovモデルでロボットの動作角の時系列をモデル化し、Forward-Backwardでベイズ学習 (動作の“単語”を知る)
- 腕の運動を使った分割結果：



分節結果

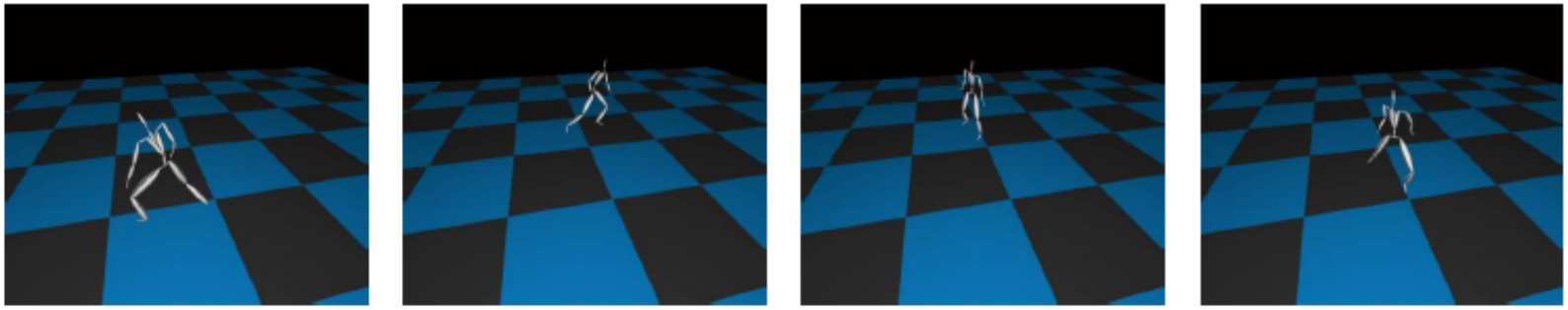
実験：分節化結果



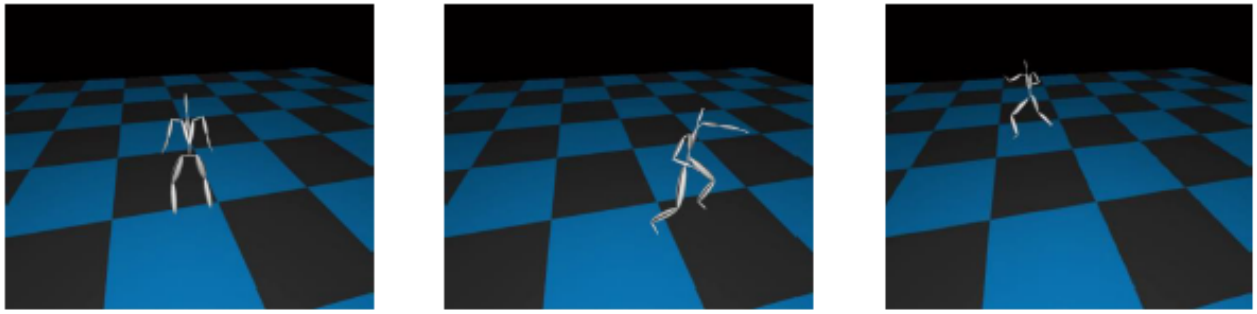
- おおむね正しい切れ目を推定できている

実験：教師なしで抽出された動作

- クラス0：右追い突き（右手の一歩踏み出してのパンチ）

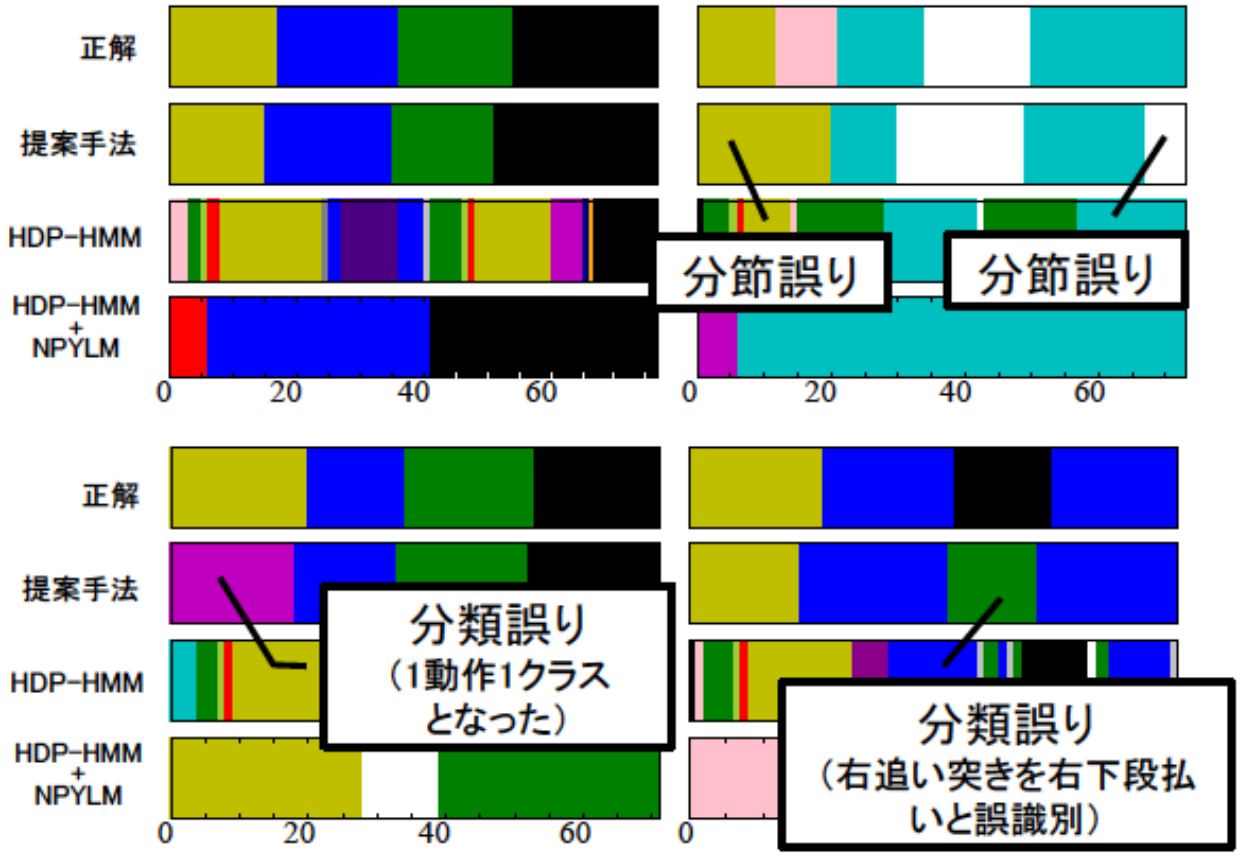


- クラス6：左下段払い（左下段のガード）



実験：比較評価

- 正解，GP-HSMM，HDP-HMM，HDP-HMM+NPYLMで比較



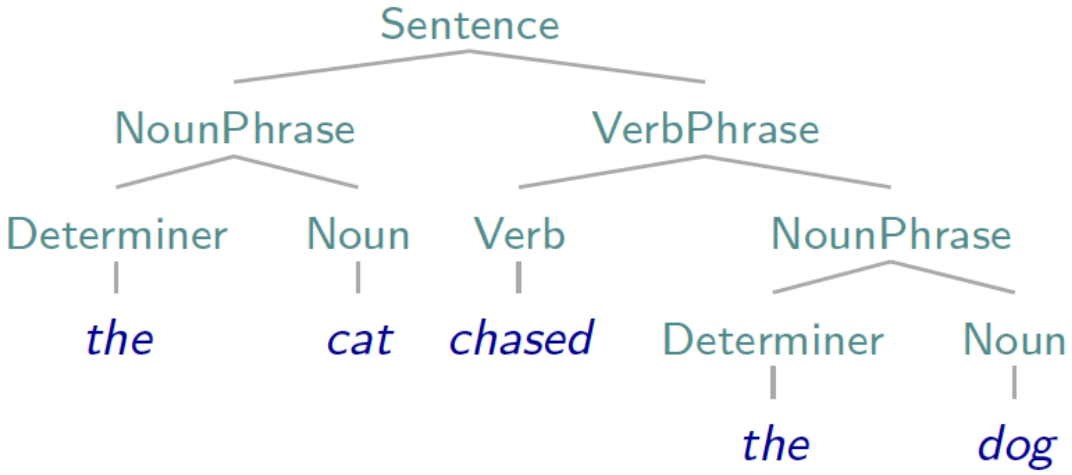
- HDP-HMMでは
  細かい分節が多い
- その影響で決まったパターン
  が出現せず，NPYLMで正し
  い分節が推定できていない
- 提案手法では，
  分節誤り2つ
  分類誤り2つ
- おおむね正しい
  分節化ができている

「分割」を超えて

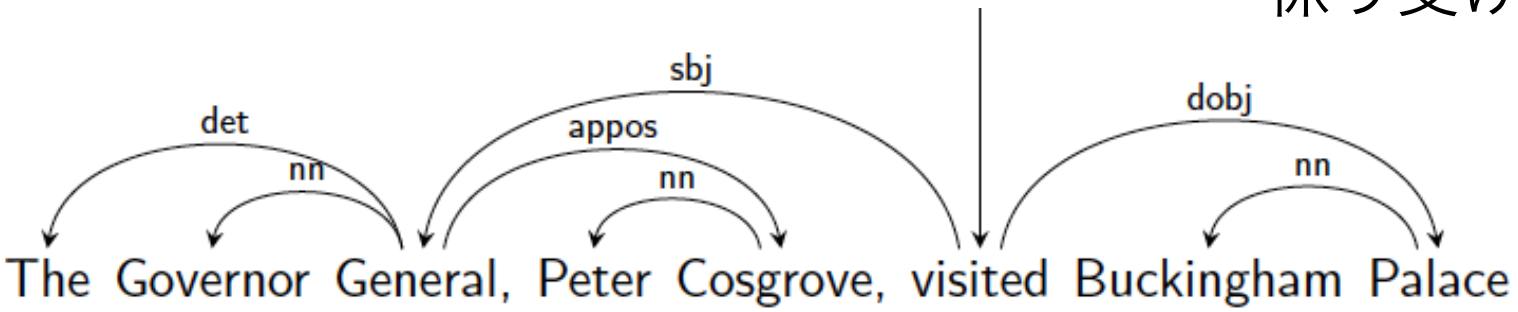
- ただ分割してタグ付けするだけから、その先へ
- 動作の「文法」と再帰的解析
- 無限再帰的状態 (無限木構造HMM: 持橋2016)

動作の「文法」

- 言語には、構文構造がある



句構造 (PCFG)



係り受け構造

統計的には...

- 分割モデルは、分割を表す潜在変数zを推定

$s =$	彼	女	の	言	っ	た	言	葉	は	$\cdots$
$\mathbf{z} =$	0	1	1	1	0	1	0	1	1	$\cdots$
$\mathbf{w} =$	彼女		の	言	った		言葉		は	$\cdots$

$$p(\mathbf{w}) = \sum_{\mathbf{z}} p(\mathbf{w}, \mathbf{z}) = \sum_{\mathbf{z}} p(\mathbf{w}|\mathbf{z})p(\mathbf{z})$$

- zは何でもよい!
  - 木構造、係り受け、感情、etc.

$$p(\mathbf{w}) = \sum_{\mathbf{t}} p(\mathbf{w}, \mathbf{t}) = \sum_{\mathbf{t}} p(\mathbf{w}|\mathbf{t})p(\mathbf{t})$$

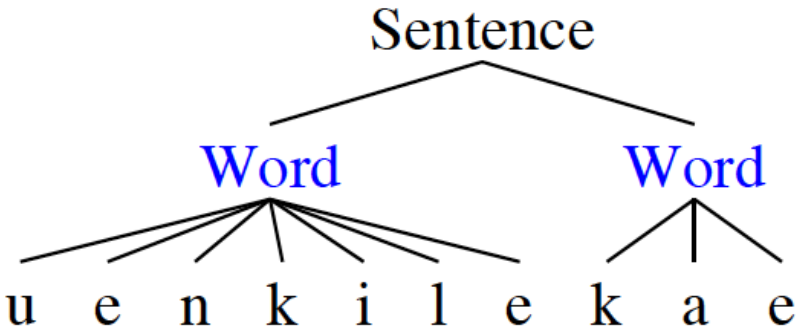
例: 文法構造による分割 (Johnson 2008)

- 文の文字列が、次のような規則で再帰的に生成されたと仮定:
  - Sentence→Colloc+
  - Colloc→Word+
  - Word→SyllableIF
  - Word→SyllableI (Syllable) (Syllable) SyllableF
  - Syllable→Onset Rhyme
  - SyllableIF→OnsetI RhymeF …

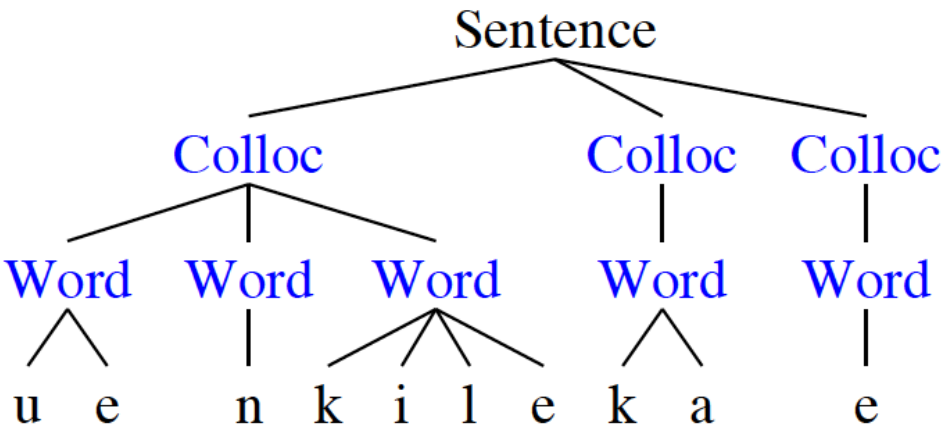
セソト語の単語と文の解析

- Sesotho: レソト(南アフリカ)の公用語, バントゥー語系

簡単な文法の場合



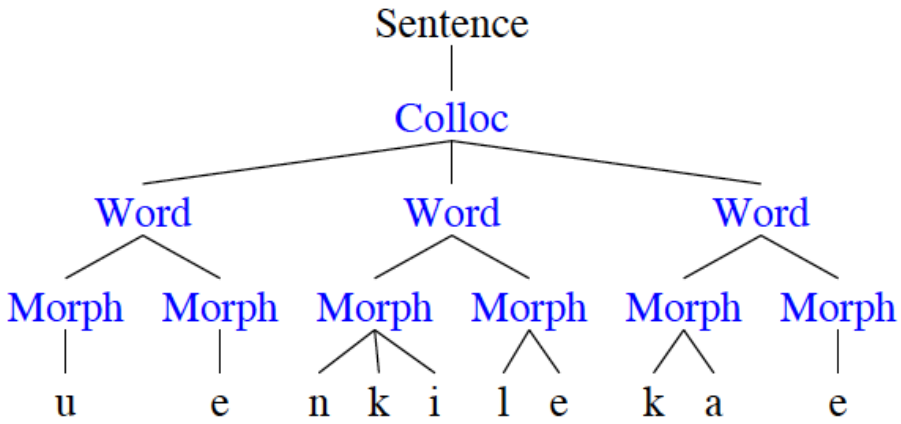
もう少し複雑な文法の場合



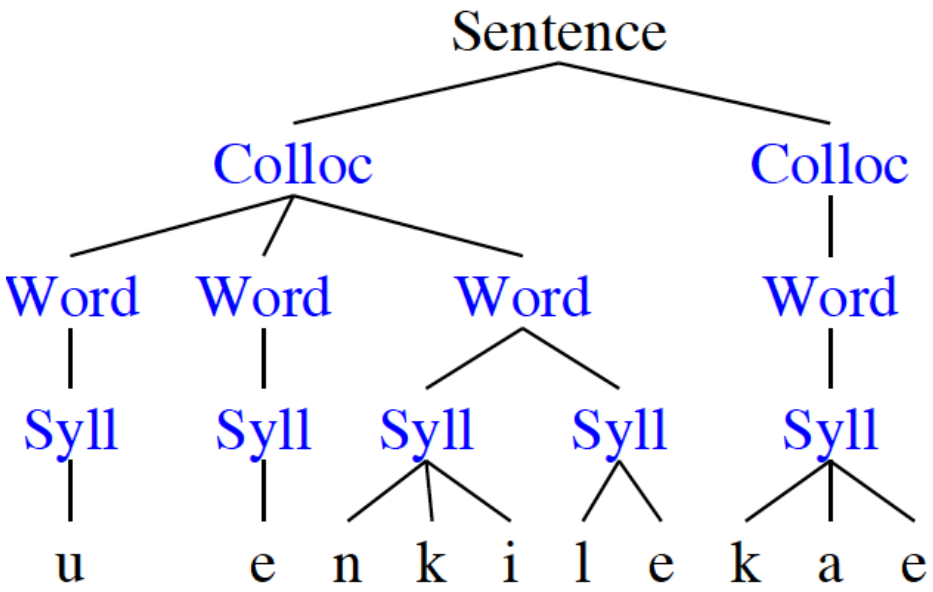
セソト語の単語と文の解析

- Sesotho: レソト(南アフリカ)の公用語, バントゥー語系

形態素の導入



単語間の句構造の導入

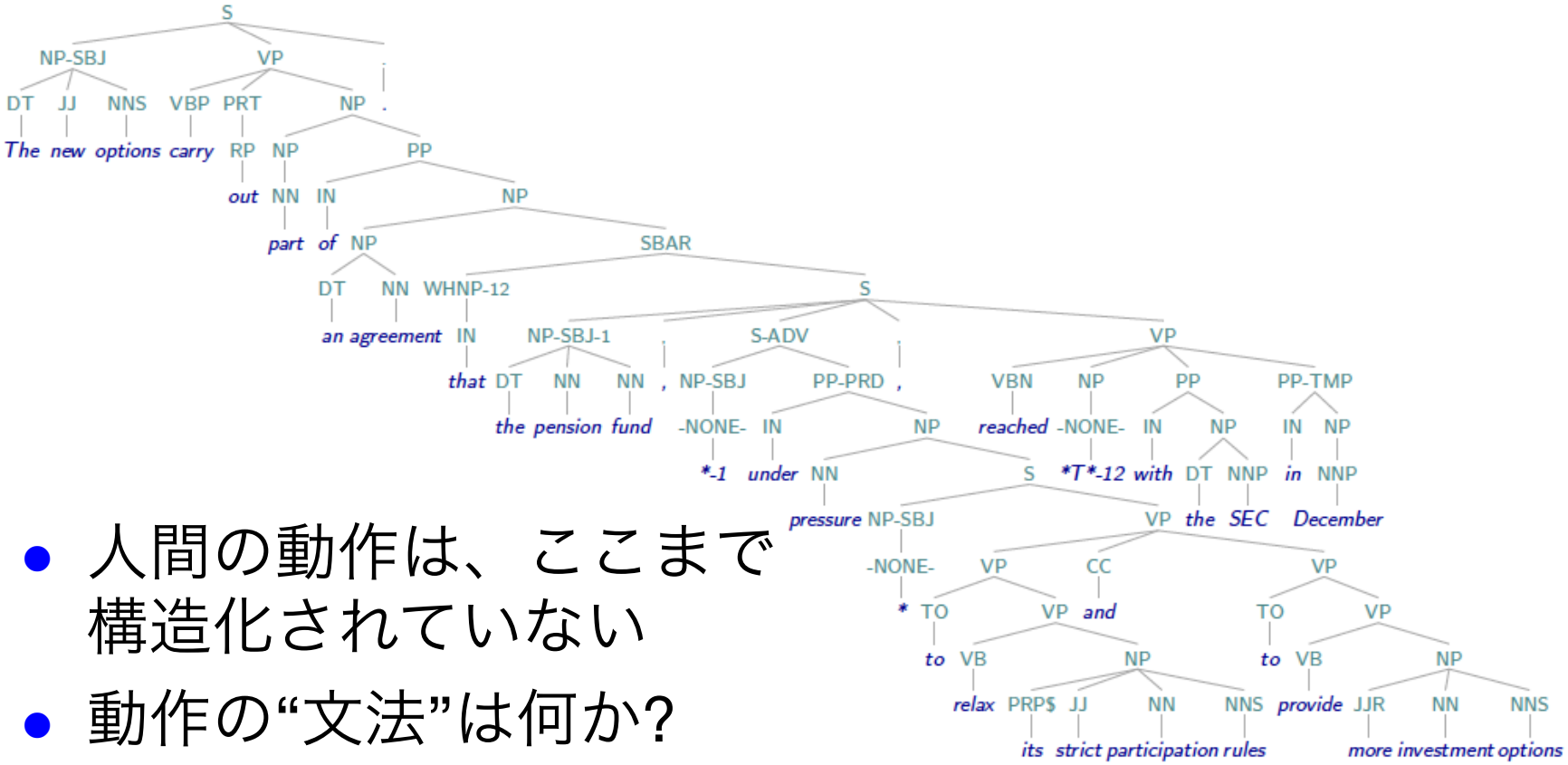


[参考] 文法構造による分割 (Johnson&Gold-water, NAACL 2009)

- 子供の発話コーパス (Bernstein-Ratnerコーパス) に対して、正解率88%程度で最高精度
  lUkD*z6b7wlThIzh&t→lUk D*z 6 b7 wIT hIz h&t
  (look there's a boy with his hat)
  - ただし、平均9.79文字/文の短い文/音素列

動作と言語

- 実際の自然言語文の解析



- 人間の動作は、ここまで構造化されていない
- 動作の“文法”は何か?

動作とHMM

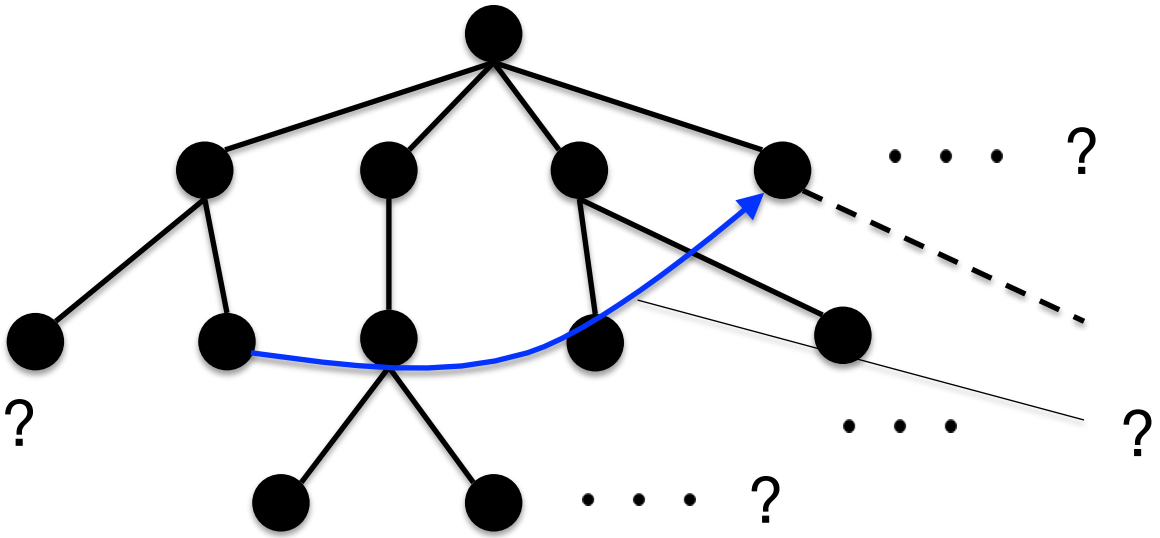
- 言語やロボティクスの解析の目標の一つは、系列を分解して“状態”を与えること
  - 言語：名詞、動詞、形容動詞、・・・・
  - ロボティクス：“走る”, “投げる”, “振り向く”, ・・・・
- 単純な状態を与えるだけで十分か?
  - 急いで走る
  - なめらかに歩く
  - さっと振り向く
  - 名詞ー固有名詞ー地名
- 状態は本来、構造化されているべき！

動作とHMM (2)

- 単純に、状態数を増やせばよい……？
  - 状態4：歩く
  - 状態23：ゆっくり歩く
  - 状態51：急いで歩く

- 普通のHMMでは、状態が完全にバラバラ！
  - 状態は、“歩く一急いで歩く”, “歩く-ゆっくり歩く”のように階層構造をしているべき
  - 自然言語では、階層的な品詞 (“動詞-他動詞-抽象”)

階層的隠れ状態の学習



- 問題:
  - 各分岐の数を何個にすればよいのか? (無限の選択)
  - どの深さまで木を考えればよいのか? (指数爆発)
  - ノード間の遷移確率をどう考えればよいのか?

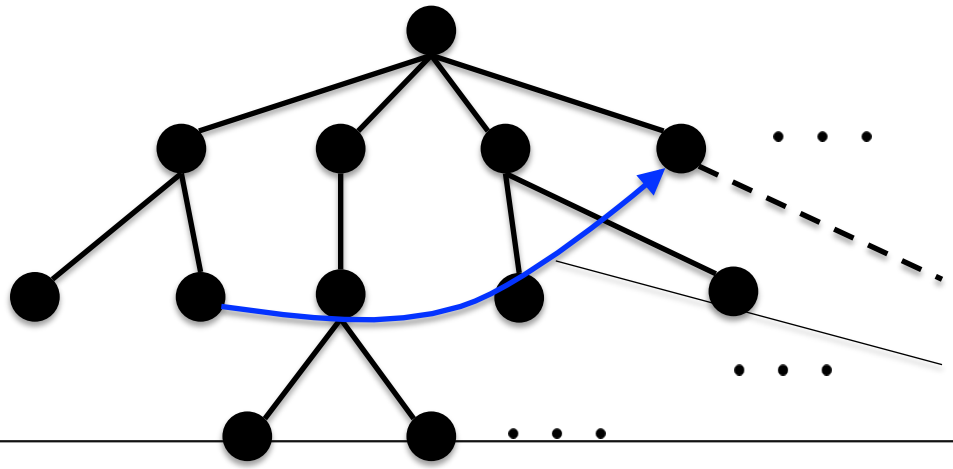
  ➡ ナイーブな方法では不可能！

HMMと状態遷移

- 通常のHMMの状態遷移: 行列で書ける

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

- 木構造間の状態遷移: どうやって定義する?

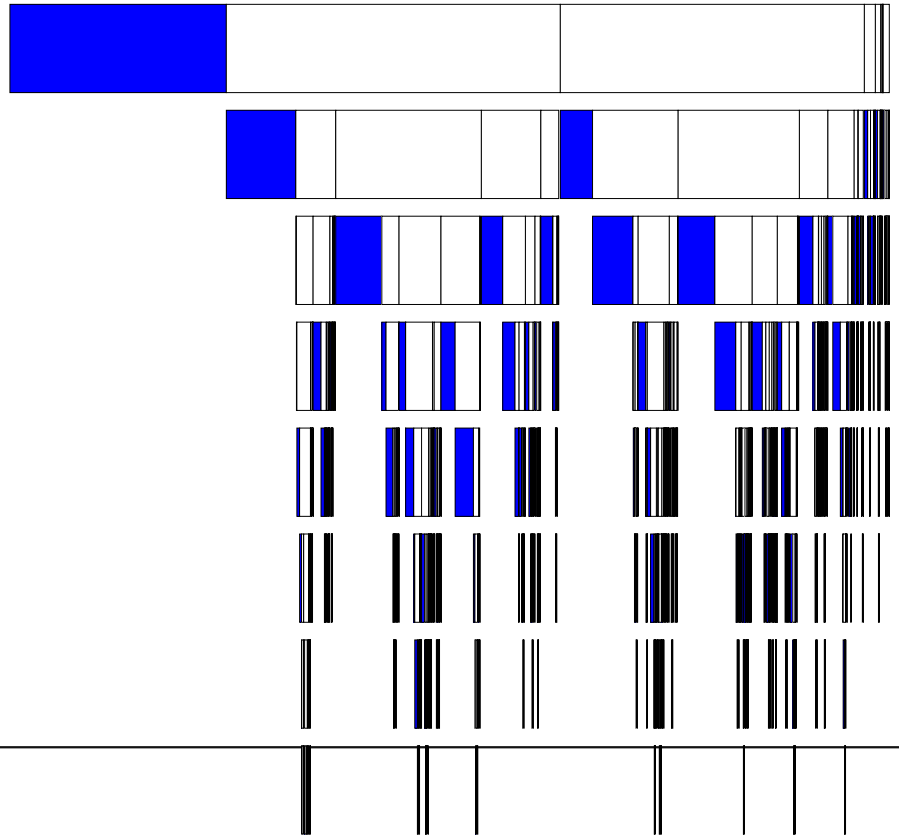


しかも、木構造の
分岐と深さは無限大

無限木構造: 木構造Stick-breaking過程

(Adams+ 2010)

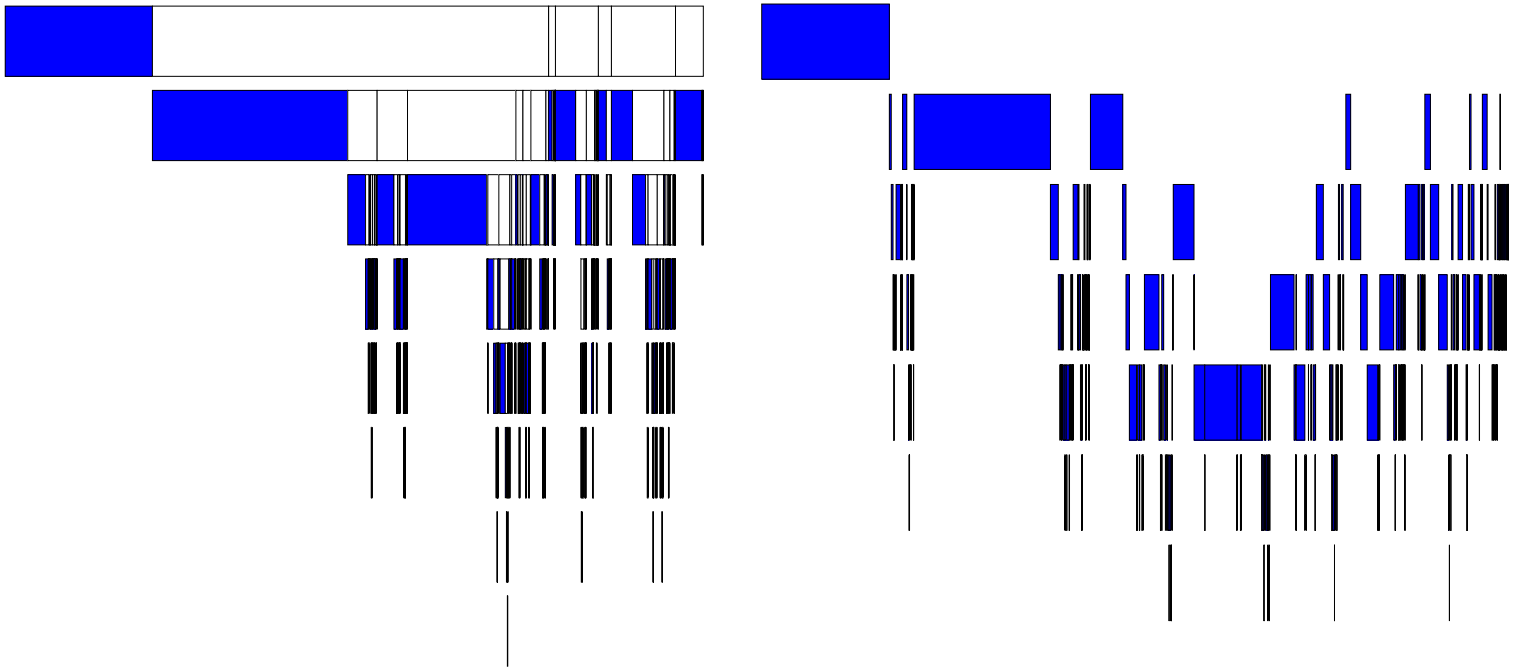
- ディリクレ過程＝Stick-breaking過程の階層化
  - “途中で止まる確率”も導入した、無限木構造の統計的生成モデル



長さ1の棒を階層的に折ることで,
和が1で、かつ
木構造を持つ
無限次元離散分布
を生成できる

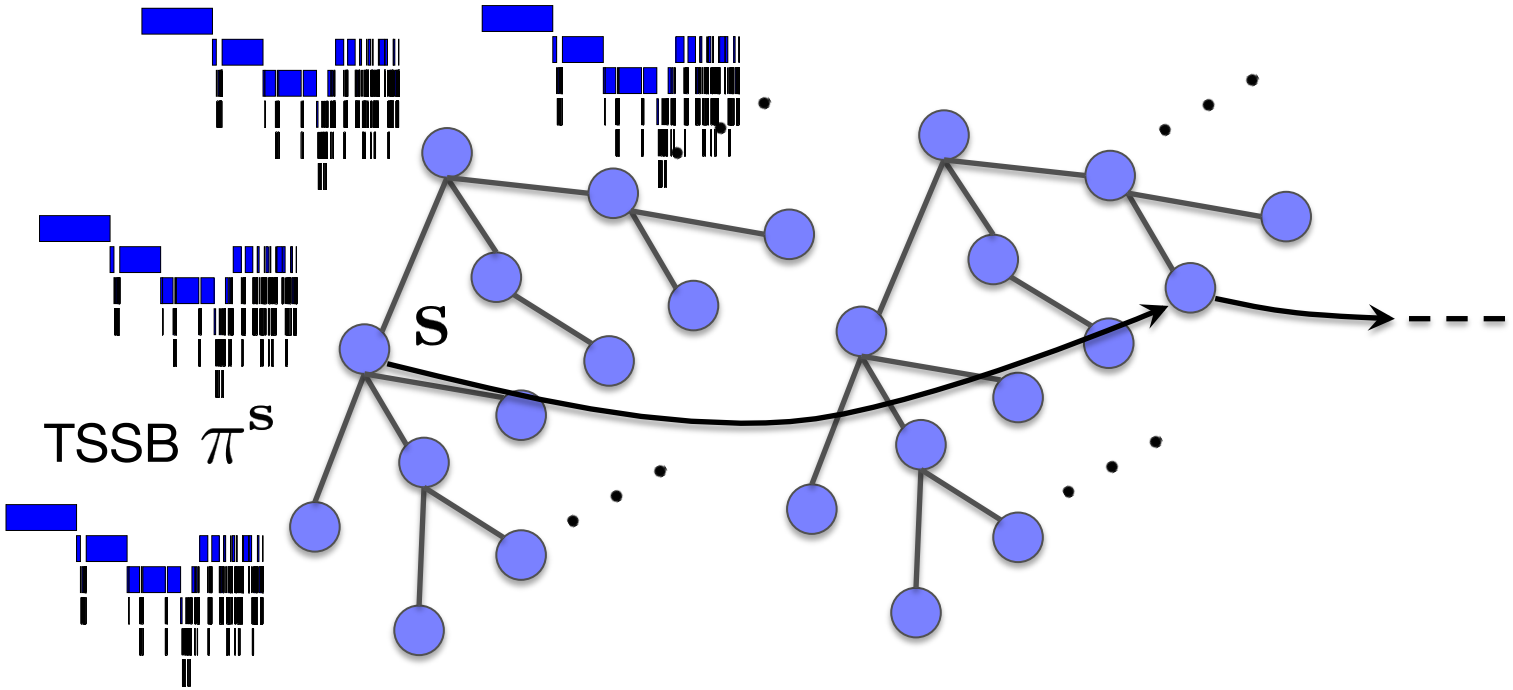
TSSBの構成

- $\pi \sim \mathrm{TSSB}(\eta, \gamma)$ から実際に生成したサンプル



TSSB(1,1,1)　　TSSB(1,5,0.5)

無限木構造上の状態遷移モデル

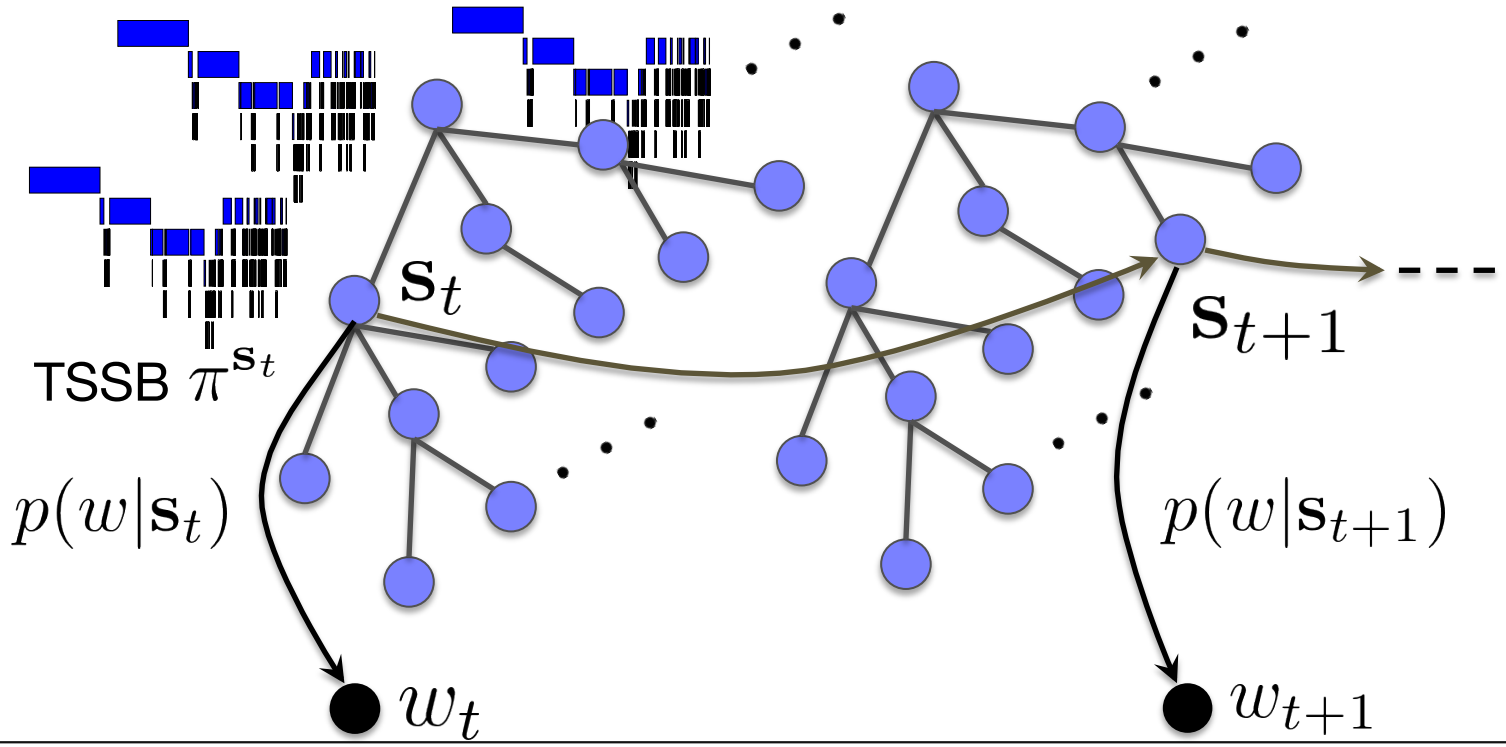


- 各ノード s が、次の状態への確率分布(TSSB) π^s を持っている
  - かつ、親子の π^s は似ているはず

無限木構造HMM (iTHMM)

(SIGNL226, 2016)

- 各ノードの持つ無限状態遷移$\pi^{\mathbf{s}}$は、それ自体親子関係
  → Hierarchical TSSB (TSSBからTSSBを生成)
- HTSSB-HMM = Infinite Tree HMM (iTHMM)



実験 (1)

- 教師なし学習: “Alice in Wonderland”, 学習1200文, テスト231文

[]			[0 0]		
next	13	0.0027	don’t	50	0.0650
one	9	0.0004	could	43	0.0563
that	8	0.0017	are	31	0.0404
mind	7	0.0004	can	30	0.0391
two	7	0.0004	would	28	0.0358
indeed	6	0.0004	must	27	0.0351
round	6	0.0004	might	24	0.0311
bill	6	0.0004	should	23	0.0298

実験 (1)

- 教師なし学習: “Alice in Wonderland”, 学習1200文, テスト231文

[4]			[4 0]		
mock	52	0.0413	voice	33	0.0542
queen	49	0.0389	way	29	0.0495
gryphon	48	0.0381	tone	26	0.0431
hatter	34	0.0263	thing	19	0.0313
mouse	33	0.0261	side	13	0.0202
duchess	29	0.0228	bit	13	0.0211
caterpillar	27	0.0212	face	13	0.0211
cat	25	0.0196	cat	12	0.0208

実験 (2)

- 半教師あり学習: 京大コーパスから10000文の品詞を教師ありデータとして固定、37400文をサンプル

[3 1]		
ついて	231	0.2009
OOV	92	0.0838
よって	73	0.0632
とって	64	0.0554
対し	63	0.0545
対して	56	0.0484
より	31	0.0266
して	25	0.0216

[3 1 6]		
よる	297	0.5674
対する	97	0.1852
関する	41	0.0781
おける	17	0.0323
基づく	17	0.0323
かかわる	12	0.0227
伴う	10	0.0189
OOV	9	0.0171

連続観測値への拡張

- 簡単のため言語の離散観測値を考えていたが、ITHMMはHMM自体の本質的な拡張であり、連続値へも自然に拡張できる
  - 出力確率分布が Gaussian で、親子関係を持つ
- 現在研究中
- 適切なデータを募集しています!

Conclusion

- 時系列の分割問題は、統計的にはSemi-Markovモデルとしてとらえられる
  - 出力が文字列: 教師なし形態素解析 (持橋+ 2009)
  - 出力がガウス過程: 動作の教師なし分割 (中村+2016)
- 分割＝形態素解析は、自然言語処理ではあくまで最初のステップ
- 分割を超えて：
  - 再帰的な分割：動作の「構文解析」
  - 再帰的な状態：潜在状態の無限木構造化
  - ロボティクスに特有の連続的な拡張？